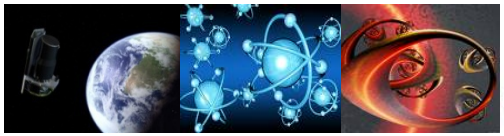
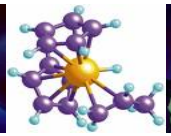



Balancing between Content Standards and Local Requirements for Scientific Metadata

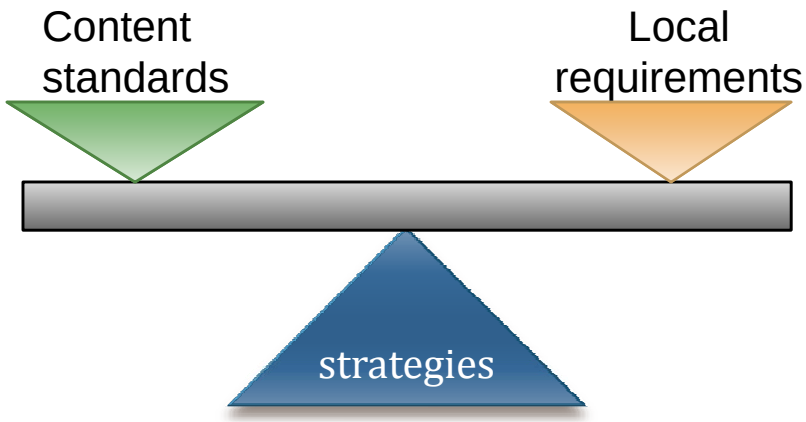
Jian Qin

School of Information Studies
Syracuse University

Presentation at Cornell University Library
September 19, 2008

Agenda

- Overview of content standards for scientific metadata
- Levels of data processing and their effects on scientific metadata
- Balancing between content standards and local requirements



Major metadata content standards

Biological sciences

Biological Data Profile

Darwin Core (DwC)

Ecological Metadata Language (EML)

Shoreline Metadata Profile

Geospatial

Georeferencing elements

FGDC CSDGM

ISO 19115: 2003 *Geographic information—Metadata.*

Climate

NetCDF Climate and Forecast (CF) Metadata Conventions

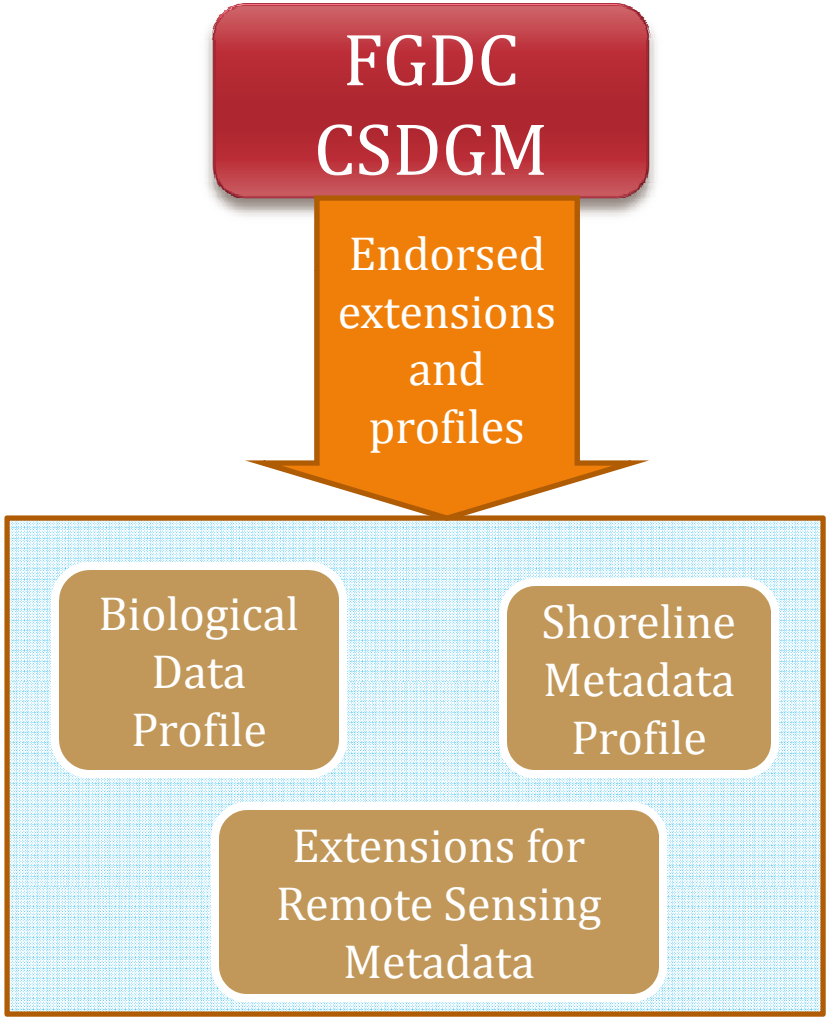
Georeferencing elements

Astronomy

Astronomy Visualization Metadata Standard

Metadata for datasets

- Provide information for dataset
  - Identification
  - Extent
  - Quality
  - Spatial and temporal schema
  - Spatial reference, and
  - Distribution



Inside the content standards: ISO 19115

- Goals:
  - Characterize geographic information
  - Facilitate geo info organization and management
  - Informing users of basic characteristics of data
  - Enable locating and access to data

Metadata entity set information	Content information
Identification information	Portrayal catalogue information
Constraint information	Distribution information
Data quality information	Metadata extension information
Maintenance information	Application schema information
Spatial representation information	Extent information
Reference system information	Citation and responsible party information

Core metadata for geographic datasets: ISO 19115

Mandatory elements:

- Abstract describing the dataset
- Dataset language
- Dataset reference date
- Dataset title
- Dataset topic category
- Metadata date stamp
- Metadata point of contact

M= Mandatory elements
C = Mandatory under certain conditions.
O = Optional elements

Conditional and Optional elements:

- Additional extent information for the dataset (vertical and temporal) (O)
- Dataset character set (C)
- Dataset responsible party (O)
- Distribution format (O)
- Geographic location of the dataset (C)
- Lineage (O)
- Metadata file identifier (O)
- Metadata standard name (O)
- Metadata standard version (O)
- Metadata language (C)
- Metadata character set (C)
- On-line resource (O)
- Reference system (O)
- Spatial representation type (O)
- Spatial resolution of the dataset (O)

Reasons for the core metadata

- Need to answer basic questions about datasets:
  - Does a dataset on a specific topic exist ('what')?
  - For a specific place ('where')?
  - For a specific date or period ('when')?
  - A point of contact to learn more about or order the dataset ('who')?
- Increase interoperability
- Allow users to understand without ambiguity the geographic data and the related metadata provided by either the producer or the distributor

ISO 19115 Geographic information – Metadata. First edition. Geneva, Switzerland: ISO, 2003. p. 15

What does it mean to scientific metadata?

- Application profiles to be developed based on ISO 19115
  - By country
  - By scientific discipline/field
  - By application or service
  - By data theme
- All application profiles are required to include the core elements
- Extensions should follow rules specified in the standard

Types of extensions

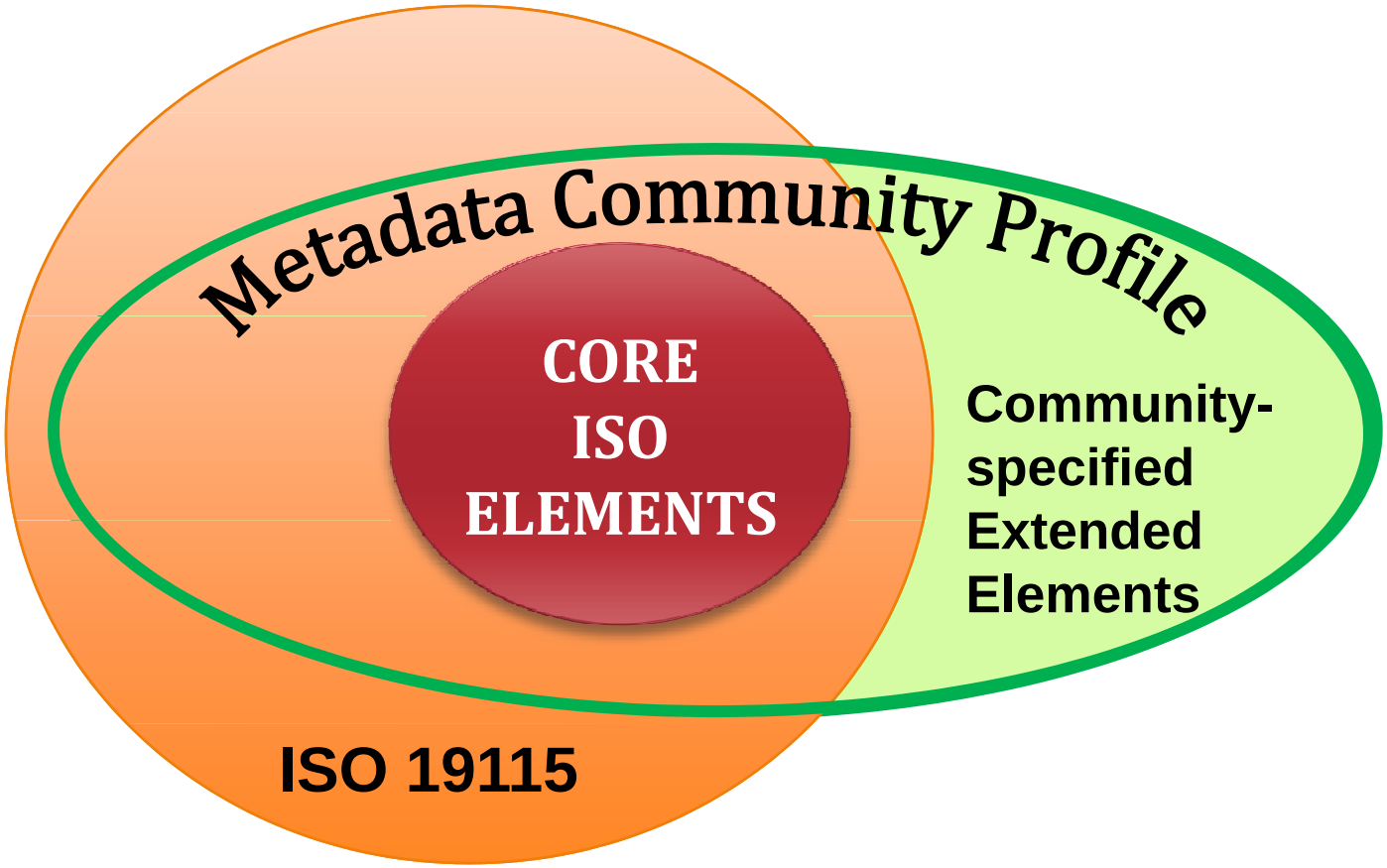
- Adding a *new metadata section*
- Creating a *new metadata codelist* to replace existing "free text" list
- Creating *new metadata codelist elements*
- Adding a *new metadata element*
- Adding a *new metadata entity*
- Imposing a *more stringent obligation* on an existing metadata element
- Imposing a *more restrictive domain* on an existing metadata element

Rules for creating an extension

- Extended metadata elements shall not be used to change the name, definition or data type of an existing element
- Extended metadata may be defined as entities and may include extended and existing metadata elements as components

ISO 19115 Geographic information – Metadata. First edition. Geneva, Switzerland: ISO, 2003. pp. 105-106.

ISO 19115 community profiles



From: FGDC. (2008). North American Profile Development for ISO 19115 Geospatial Metadata. http://www.fgdc.gov/training/nsdi-training-program/materials/ISONAPDevelopment_20080331.ppt

LEVELS OF DATA PROCESSING AND THEIR EFFECTS ON SCIENTIFIC METADATA

Levels of data processing

Data level	NASA's definition of data processing levels
Level 0	Reconstructed unprocessed instrument data at full resolutions.
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, but not applied to the Level 0 data.
Level 1B	Level 1A data that has been processed to sensor units. Not all instruments will have a Level 1B equivalent.
Level 2	Derived environmental variables (e.g., ocean wave height, soil moisture, ice concentration) at the same resolution and location as the Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency properties
Level 4	Model output or results from analyses of lower-level data

Bose, R. & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys*, 37(1), 1-28.

Scientific data formats

Scientific data formats

Common data formats
Image formats
Matrix formats
Microarray file formats
Communication protocols

Metaformats

DSV CSV XML

Data structures

Tuple Set List Array Tree

Physical data

Bits···Bytes···characters···strings

Data model

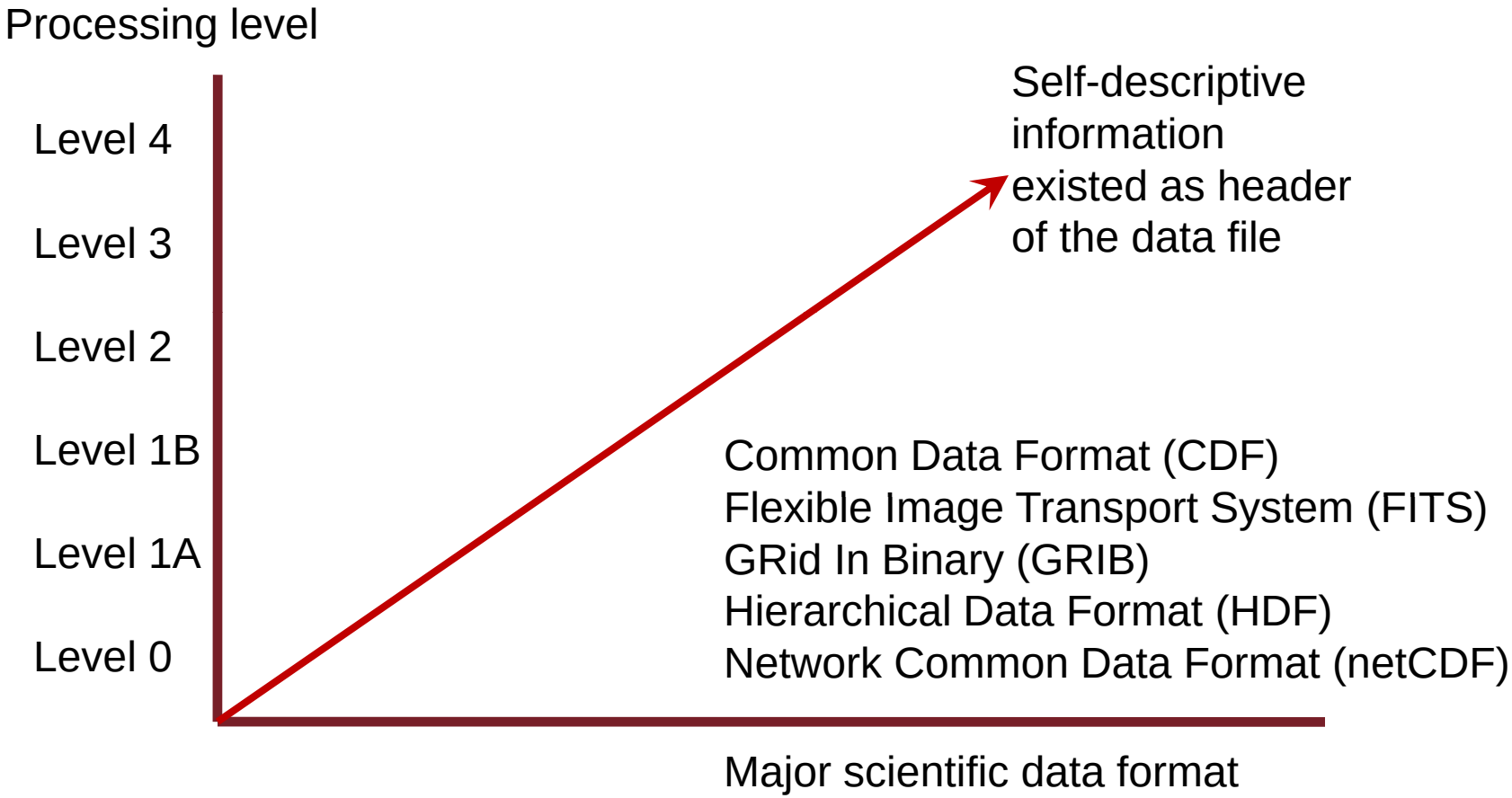
Hierarchical

Relational

Object-oriented

Network

Metadata embedded in data products



The concept of lineage

- Lineage: information about the events or source data used in constructing the data specified by the scope
  - Events or transformation in the life of a dataset
  - Source data used in creating the data
  - Process step
  - Date and time over which the process occurred
  - Spatial reference system used by the source data
  - Published references for the source data

Lineage elements in ISO 19115

DQ_DataQuality → LI_Lineage

Either LI_Source or LI_ProcessStep must be documented

LI_Source

+description
+scaleDenominator
+sourceReferenceSystem
+sourceCitation
+sourceExtent

Either description or sourceExtent must be documented.

LI_ProcessStep

+description
+rationale
+dateTime
+processor

Lineage metadata example

Lineage:
Source_Information:
Source_Citation:
Citation_Information:
Originator: U.S. Geological Survey
Publication_Date: 198811
Title: National High Altitude Program (NHAP)
Geospatial_Data_Presentation_Form: remote-sensing image
Source_Scale_Denominator: 65000
Type_of_Source_Media: black and white aerial photograph film transparency
Source_Time_Period_of_Content:
Time_Period_Information:
Single_Date/Time:
Calendar_Date: 198811
Source_Currentness_Reference: source photography date
Source_Citation_Abbreviation: NWI1a
Source_Contribution:
aerial photo from which wetlands spatial and attribute information are interpreted

Source: http://together.net/~bspatial/duck/data/pajrivsv.html#Data_Quality_Information

Lineage metadata example (cont'd)

Process_Step:
Process_Description:
NWI maps are compiled through manual photo interpretation of NHAP or NAPP aerial photography, supplemented by soil surveys and field checking of wetland photo signatures. Delineated wetland boundaries are manually transferred from interpreted photos to USGS 7.5 minute topographic quadrangle maps and then manually labeled. Quality control steps occur throughout the photo interpretation, map compilation, and map reproduction processes.
Source_Used_Citation_Abbreviation: NWI1a
Source_Used_Citation_Abbreviation: NWI1b
Source_Used_Citation_Abbreviation: NWI2
Process_Date: 1992
Source_Produced_Citation_Abbreviation: NWI3

Source: http://together.net/~bspatial/duck/data/pajrivsv.html#Data_Quality_Information

Data collections

- Research collections: generated by investigator or team
- Resource collections: created by a community of investigators in a domain
  - often developed with community-level standards
- Reference collections: created by large segments of science and engineering community
  - conform to robust, well-established and comprehensive standards

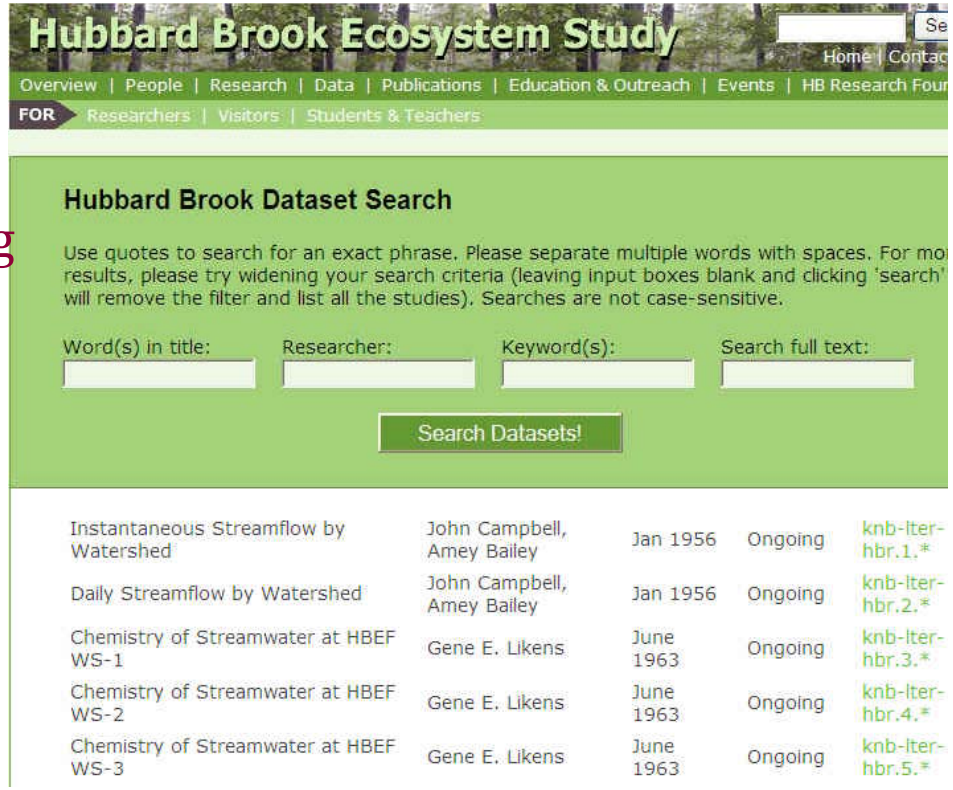
NSF. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

Research collections

- Limited processing or long-term management
- Not conformed to any data standards
- Varying sizes and formats of data files
- Low level of processing, lack of plan for data products
- Low awareness of metadata standards and data management issues

Resource collections

- Example: Hubbard Brook Ecosystem Study (http://www.hubbardbrook.org)
  - One of the regional sites in the Long term Ecological Research Network (LTER)
  - Community of a science domain
  - Community of investigators from around the country on ecosystem study
  - *Ecological Metadata Language (EML)*, a community-level standard
  - Cataloged, searchable dataset collections



Implications to metadata

Processing levels

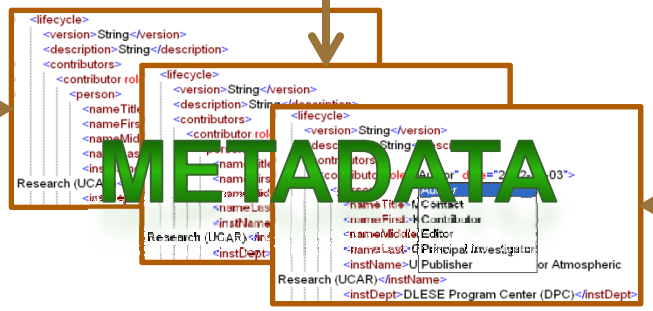
Lineage vital to assessing data quality

Data formats

Some formats contain self-descriptive metadata

Data collections

Metadata standards need to be adjusted for local description needs



How can we generate good quality metadata for scientific data with the least effort and resource?

BALANCING BETWEEN CONTENT STANDARDS AND LOCAL REQUIREMENTS

The paradox of standards and local requirements

Standards	Local requirements
• Large numbers of elements and complex structures • Focus on describing data products (datasets, data series, collections) • Little guidance on content recording • Not concerned about implementation	• Discipline-, community-, and application-bound • Focus on data management at all stages of projects and processing • Strong emphasis on best practices for content recording • Concerned about implementation in terms of costs, scalability, ease of use, etc.

Strategy: Know thy data

Which processing level?

Documentation (user guide, readme, etc.) may contain lineage information. Also help determine whether a metadata record should be created for what scope of the data

Data collections

What format?

Some format has self-descriptive metadata and can be extracted by computer program

"little science,"
"big science"

"Little science" data is more likely to be the research collection type while "big science" data tends to be the resource or reference collection type.

Strategy: adapting standards to local needs

- Application profiles at:
  - Community level
  - Discipline/fields/domain level
  - Collection level
  - Cross-community/domain/collection level
- What do they mean to metadata design?
  - Types of extensions necessary
  - Core elements from standards vs. local cores
  - Modeling of schema encodings
  - Tools for content recording
  - Local metadata registries
  - Best practice guidelines

- Abstract describing the dataset (M)
- Dataset language (M)
- Dataset reference date (M)
- Dataset title (M)
- Dataset topic category (M)
- Metadata date stamp (M)
- Metadata point of contact (M)
- Additional extent information for the dataset (vertical and temporal) (O)
- Dataset character set (C)
- Dataset responsible party (O)
- Distribution format (O)
- Geographic location of the dataset (C)
- Lineage (O)
- Metadata file identifier (O)
- Metadata standard name (O)
- Metadata standard version (O)
- Metadata language (C)
- Metadata character set (C)
- On-line resource (O)
- Reference system (O)
- Spatial representation type (O)
- Spatial resolution of the dataset (O)

Balancing between standards and local needs: cases

- For discovering:
  - Biodiversity data:

    http://knb.ecoinformatics.org/knb/metacat
- For analysis:
  - Climate dataset:

    http://www.cgd.ucar.edu/vemap/v2climate.html

Strategy: The outgoing data librarianship

- Data is neither owned nor stored in the library
- Scientists are not aware that librarians can help
- Sell data librarianship to scientists
- What librarians can contribute:
  - Help research teams assess data management needs
  - Design of data management plans including metadata applications
  - Help implement the plans
  - Manage ongoing changes in data management
  - Provide science data literacy training for future science workforce

DataStaR

Cornell University
Cornell Institute for Social and Economic Research
About Us | Computing | Data | Trai
CISER Data Archive

CUGIR Cornell University Geospatial Information Repository
Home | About | Browse | Map browse | Basket | Help

Strategy: Collaborative data librarianship

Institution

Financial and policy support

IT and other groups

Data librarian

Community

Science domain

Data content idiosyncrasies

User requirements

Evolving and interconnecting –

Institutional repository

Community repository

National repository

International repository

Summary

- Scientific metadata standards are defined to describe data products with all aspects
- Local applications adopt standards with constraints of science domains, community needs, and resources available for implementation
- Balancing between standards and local needs implicates careful design and implementation of metadata artifacts
- Data librarianship is outgoing and collaborative

The Scientific Data Literacy Project

Jian Qin (PI)

Ruth Small (co-PI)

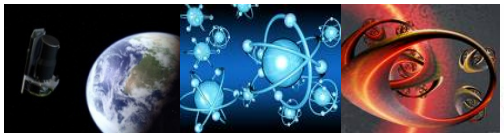
John D'Ignazio (Research Assistant)

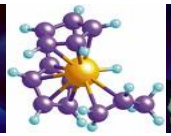
Goal:

1) Create a Scientific Data Literacy (SDL) course

2) Prepare students majoring in science and technology for a career in scientific data management

- What the project does:
  - *Assessing the needs for scientific data literacy* education through environmental scanning and surveying science and technology faculty members.
  - *Creating learning strategies, techniques, and materials* on scientific data and their lifecycle.
  - *Evaluating the effectiveness* of learning materials and pedagogy through outcome-based evidence.
  - *Generalizing and communicating the lessons learned* for larger scale implementation of the course curriculum throughout undergraduate institutions.




School of Information Studies
Syracuse University

Thank you!

Questions?